

1.

The screenshot shows the Coursera interface for the course '機器學習基石下 (Machine Learning Foundations)---Algorithmic Foundations'. The sidebar on the left lists 'Three Learning Principles' with video links: 'Occam's Razor' (10 min), 'Sampling Bias' (11 min), 'Data Snooping' (12 min), 'Power of Three' (8 min), and '測驗: 作業四' (20 questions). The main content area displays '測驗 作業四' (Assignment 4) with a score of 100.00%. A button '再次參加' (Retake) is visible. The top navigation bar includes the Coursera logo, a search bar, and a user profile icon.

2.

$$\begin{aligned}\nabla E_{aug}(\mathbf{w}) &= \nabla E_{in}(\mathbf{w}) + \nabla \left(\frac{\lambda}{N} \mathbf{w}^T \mathbf{w} \right) \\ &= \nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N} \mathbf{w}\end{aligned}$$

So, the update rule is:

$$\begin{aligned}\mathbf{w}(t+1) &\leftarrow \mathbf{w}(t) - \eta \nabla E_{aug}(\mathbf{w}(t)) = \mathbf{w}(t) - \eta \left(\nabla E_{in}(\mathbf{w}(t)) + \frac{2\lambda}{N} \mathbf{w}(t) \right) \\ &= \left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}(t) - \eta \nabla E_{in}(\mathbf{w}(t))\end{aligned}$$

3.

$$\text{Let } A = (-1, 0), B = (\rho, 1), C = (1, 0)$$

$$E_{loo} = \frac{1}{3} (\text{error}(\text{leave } B \text{ out}) + \text{error}(\text{leave } A \text{ out}) + \text{error}(\text{leave } C \text{ out}))$$

Let h_{xy} be the corresponding linear hypothesis which have the lowest square error on point x and y, which is the line constructed by x and y (the square error is zero).

$$h_{AC}(x) = 0, \text{error}(\text{leave } B \text{ out}) = 1$$

$$h_{BC}(x) = \frac{1}{\rho - 1} (x - 1), \text{error}(\text{leave } A \text{ out}) = (h_{BC}(-1) - 0)^2 = \frac{4}{(\rho - 1)^2}$$

$$h_{AB}(x) = \frac{1}{\rho + 1} (x + 1), \text{error}(\text{leave } C \text{ out}) = (h_{AB}(1) - 0)^2 = \frac{4}{(\rho + 1)^2}$$

$$\text{So, } E_{loo}(\rho) = \frac{1}{3} \left(1 + \frac{4}{(\rho - 1)^2} + \frac{4}{(\rho + 1)^2} \right)$$

4.

$$\text{Let } E_{aug}(\mathbf{w}) = \frac{\lambda}{N} \|\mathbf{w}\|^2 + E_{in}(\mathbf{w}) = \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

$$\nabla E_{aug}(\mathbf{w}) = \frac{2\lambda}{N} \mathbf{w} + \nabla E_{in}(\mathbf{w}) = \frac{2\lambda}{N} \mathbf{w} + \frac{2}{N} (\mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y})$$

$$\text{SGD: } \nabla E_{aug}(\mathbf{w}) = \frac{2\lambda}{N} \mathbf{w} + \nabla E_{in}(\mathbf{w}) \approx \frac{2\lambda}{N} \mathbf{w} + 2(\mathbf{w}^T \mathbf{x}_n \mathbf{x}_n - y_n \mathbf{x}_n)$$

(use point (\mathbf{x}_n, y_n) to approximate the true gradient)

So the update rule is:

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \eta \nabla E_{aug}(\mathbf{w}_t) \\ &= \mathbf{w}_t - \eta \left(\frac{2\lambda}{N} \mathbf{w}_t + 2(\mathbf{w}_t^T \mathbf{x}_n \mathbf{x}_n - y_n \mathbf{x}_n) \right) \\ &= \left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}_t - \eta \left(2(\mathbf{w}_t^T \mathbf{x}_n \mathbf{x}_n - y_n \mathbf{x}_n) \right) \end{aligned}$$

The update rule in Q3 looks like this:

$$\mathbf{w}_{t+1} = \left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}_t - \eta \nabla E_{in}(\mathbf{w}_t)$$

We can observe that $\left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}_t - \eta \left(2(\mathbf{w}_t^T \mathbf{x}_n \mathbf{x}_n - y_n \mathbf{x}_n) \right) \approx \left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}_t - \eta \nabla E_{in}(\mathbf{w}_t)$

So, the two update rules are probably approximately same, meaning that when we add virtual example $\tilde{\mathbf{X}} = \sqrt{\lambda} \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$ in the training data set and do normal SGD without regularization, it can reach the same result using GD with regularization.

5.

For target function $\sin(ax)$, $x \in [0, 2\pi]$, the squared error for $h(x) = wx$ is:

$$\begin{aligned} err(w) &= \int_0^{2\pi} (\sin(ax) - wx)^2 dx = -\frac{2wsin(2\pi a)}{a^2} - \frac{\cos(2\pi a)(\sin(2\pi a) - 8\pi w)}{2a} + \frac{8\pi^3 w^2}{3} + \pi \\ \frac{\partial err(w)}{\partial w} &= -\frac{2sin(2\pi a)}{a^2} + \frac{\cos(2\pi a)(8\pi)}{2a} + \frac{16\pi^3 w}{3} \end{aligned}$$

To solve $\min_w err(w)$, we need to solve the equation $\frac{\partial err(w)}{\partial w} = 0$. After calculation, we get:

$$w = \frac{\frac{2sin(2\pi a)}{a^2} - \frac{\cos(2\pi a)(8\pi)}{2a}}{\frac{16\pi^3}{3}} = \frac{3sin(2\pi a) - 6\pi a \cos(2\pi a)}{8\pi^3 a^2}$$

So, for each x , the level of deterministic noise is $\left| \sin(ax) - \frac{3sin(2\pi a) - 6\pi a \cos(2\pi a)}{8\pi^3 a^2} x \right|$